科研实体名称规范的关联数据模型构建*

周毅　张建勇　刘峥　刘秀敏
中国科学院文献情报中心　北京 100190

摘　要：[目的/意义]旨在研究将国家科技图书文献中心(National Science and Technology Library, NSTL)的科研实体名称规范数据发布为关联数据的难点——关联数据的数据模型。科研实体名称规范数据的数据模型研究，有助于NSTL 科研实体数据的共享、互联、质量提升，融入到互联网中，同时也为其他机构使用、发布关联数据提供模型参考。[方法/过程]首先，分析比较国内外关联数据发布项目中所采用的数据模型，发现关联数据发布项目中的数据模型主要分为以 Schema. org 为核心和多种标准词表组合两类；结合 NSTL 名称规范数据的特点，设计两种形式的关联数据模型，并从关联数据模型对名称规范数据的表达程度、模型复杂度等角度进行比较，选择较优方案；最后以 D2RQ 为工具进行实验，将 NSTL 名称规范的样例数据发布为关联数据。[结果/结论]分析发现两种方案中以 Schema. org 为核心标准词表的方案相对于多种标准词表组合的方案有较优的表达完整度、较低的模型复杂度，更易于融入互联网，因此更适合作为 NSTL 名称规范数据的关联数据模型。

关键词：科研实体　名称规范　关联数据　数据模型
分类号：G254
DOI：10.13266/j. issn. 0252 – 3116. 2020. 10. 012

1　引言

科研实体是科研活动的组成要素，主要包括科研主体(科研人员、科研机构等)、科研活动、科研条件(科研方法、科学文献、科学期刊等)、科研产出(科研成果等)[1]。规范控制的本质是实现基于概念的描述和匹配[2]，其目的是汇集同一实体的不同名称形式，并区分具有相同名称的同一实体，实现语义消歧。人员、机构、出版物、基金等科研实体信息在信息服务中发挥着重要作用。NSTL 对其多个系统和出版社数据中的科研实体信息进行规范，形成了涵盖机构、人员、基金、期刊 4 类名称规范数据的 NSTL 名称规范系统。

NSTL 的名称规范数据目前遵循的是《名称规范元数据标准》，其以模块化、最小粒度等为设计原则，使得规范对象的信息能够被深入细致地描述。但其以单条记录为描述单元，没有形成以真实世界实体为对象的描述框架或本体，不能支持数据集的重用与开放关联。同时，NSTL 名称规范数据从文献数据中析出，描述规范对象的信息有限，数据稀疏，需要通过与其他关联数据集融合以提高质量。

因此，为了使规范数据充分发挥应用价值，提高数据质量，有必要研究如何将 NSTL 名称规范数据模型转换为关联数据模型，发布为关联数据。NSTL 作为国家科技文献信息的资源保障基地、服务集成枢纽和服务发展支持中心，将其名称规范数据发布为关联数据有着重要意义：①使数据描述转向真实世界的实体及实体关系；②可供 NSTL 其他系统或第三方机构重用；③可与其他关联数据集融合，提高数据完整程度，从而提高数据质量。

构建数据模型是关联数据发布过程中的关键环节。如何使数据更好地为机器所理解，如何使数据更方便地与其他关联数据关联，都依赖于一个科学的、可持续的数据模型。本文借鉴关联规范数据发布项目中的数据模型构建经验，力图构建适用于 NSTL 名称规范的关联数据模型。通过将 NSTL 名称规范数据模型转换为关联数据模型，实现关联数据发布。

* 本文系国家科技图书文献中心(NSTL)资助项目“名称规范数据库建设”(项目编号：科 1817)研究成果之一。
作者简介：周毅(ORCID：0000 – 0002 – 1494 – 6716)，馆员，硕士；张建勇(ORCID：0000 – 0001 – 7533 – 1726)，研究馆员，硕士；刘峥(ORCID：0000 – 0002 – 2494 – 436X)，副研究馆员，博士，通讯作者，E-mail：liuz@mail. las. ac. cn；刘秀敏(ORCID：0000 – 0001 – 6014 – 9614)，馆员，硕士。
收稿日期：2019 – 11 – 25　修回日期：2020 – 02 – 19　本文起止页码：109 – 117　本文责任编辑：易飞

2 NSTL 的科研实体名称规范数据

NSTL 汇集了 NSTL 加工系统和历史系统的文献数据以及 WoS、Wiley 等 9 家出版社的文献数据，从这些文献数据中抽取机构、人员、基金、期刊等科研实体信息，建立起 NSTL 名称规范系统。名称规范系统中涵盖机构、人员、基金、期刊 4 种名称规范数据，具体包括期刊数据 3 万余条，人员数据 1 700 多万条，机构数据 350 多万条，基金数据 300 余万条。

NSTL 名称规范数据根据《名称规范元数据标准》，从文献元数据中析出。规范数据中描述各类对象的元素和属性数量见表 1。表 1 中人员有姓名、联系方式、出生与死亡日期、所属机构方面的元素和属性；机构有机构名称、联系方式、地址、历史信息等方面的元素和属性；基金有基金名称、日期、主题、介绍等方面的元素和属性；期刊有期刊名称、标识符、出版日期、期信息等方面元素和属性。这里的属性区别于本体中的属性，它是元素的组成部分，定义元素的特性。一条规范数据实际上是关于单个规范对象的一组记录，这组记录可能包含多个记录并来自多个不同的数据源，其中一条记录为优选名称，其余记录作为其他名称存在。

表 1 描述各规范对象的元素与属性

序号	规范对象	元素数量(个)	属性数量(个)
1	人员	21	15
2	机构	17	14
3	基金	15	14
4	期刊	30	11

《名称规范元数据标准》虽然深入细致地描述了规范对象，但没有形成以真实世界实体为描述对象的框架或本体，不符合关联数据发布需求。将 NSTL 名称规范数据模型转换为关联数据模型，发布为关联数据，需要关联数据模型能够充分表达已有数据，同时具备与其他关联数据集的互操作性。在构建关联数据模型的过程中，我们可以借鉴国内外关联数据发布项目的经验。

3 国内外关联数据发布项目中的数据模型

在语义网社区和开放关联数据运动的推动下，越来越多的图书馆及文化机构将自身的规范数据发布为关联数据。本文选择国内外比较有代表性的项目，分析其关联数据模型的内容和特点，选择的实践项目包括联机计算机图书馆中心（Online Computer Library Center，OCLC）的虚拟国际规范文档（Virtual International Authority File，VIAF）、德国图书馆综合规范文档（Gemeinsame Normdatei，GND）、上海图书馆人名规范库、科研人员信息平台 VIVO 等。

3.1 关联数据发布项目中的数据模型概述

OCLC 管理的 VIAF 将多个名称规范文件合并到一个单独的名称规范服务中。其目标是通过匹配和链接规范文件，在线提供规范信息。VIAF 目前已有 30 多个国家的 40 多个组织的数据[3]。VIAF 的数据模型经过了多次演变，逐渐将描述重点由概念、名称转为实体。VIAF 最初将规范对象看作概念，后来区分了不同的实体，包括地点、机构、人员、作品等。最初 VIAF 几乎全部采用简单知识组织系统（Simple Knowledge Organization System，SKOS）描述信息，随后又添加了朋友的朋友（Friend of a Friend，FOAF）、资源描述与检索（Resource Description & Access，RDA）描述。一段时间内它们同时存在，从不同的侧面对对象进行描述[4]。2014 年 VIAF 追随 Wikidata 转为以 Schema.org 为核心词汇表描述[5]。除此之外，2012 年，OCLC 以 Schema.org 及其图书馆扩展（SchemaBibEx）为基础的书目数据实验模型将世界上最大的在线联合目录 WorldCat 发布为关联数据[6]。

2010 年起，德国国家图书馆开始尝试将其规范数据发布为关联数据[7]。德国国家图书馆 GND 汇集了企业机构规范文档（Corporate Body Authority File，GKD），人名规范文档（Name Authority File，PND）、主题规范文档（Subject Headings Authority File，SWD）以及德国音乐档案馆统一标题文档（Uniform Title File，EST）的内容[8]。GND 的目标是形成一个网络兼容的规范文档，连接德语国家的图书馆和其他文化机构所拥有的各种信息资源，以语义网的方式将数据开放到图书馆界之外，向学术和文化界的各种用户开放[9]。GND 综合规范文档以 GND 本体为数据模型[7]。GND 本体中的实体大类包括人、会议或事件、法人团体、地点或地理名称、主题、作品，大类下区分了多个下位类。为确保兼容性，GND 本体保持与现有标准词表如 FOAF、RDA 等的一致性[8]。

2017 年，上海图书馆利用本体和关联数据技术，将传统的名称规范文档关联数据化，建立了一个新的人名规范库。该数据库融合上海图书馆的系谱学、珍本馆藏、档案馆和其他特殊馆藏的名称，汇集了 840 000个人的信息。人名规范数据库不仅提供关于人的基本信息，还提供关键事件、社会关系以及可以显示、搜索、汇总和分析的作品[10]。上海图书馆人名规

范库的数据模型以 IFLA-LRM 和 BIBFRAME 为基础。上海图书馆为人名规范库构建了人名规范库本体(shl-names),主要的类有人、地点、时间、事件、物理对象和资料(见图 1)。本体中复用 FOAF 描述人员类、Relation 词表表示类之间的关系,复用 PROV Ontology 描述事件[11]。

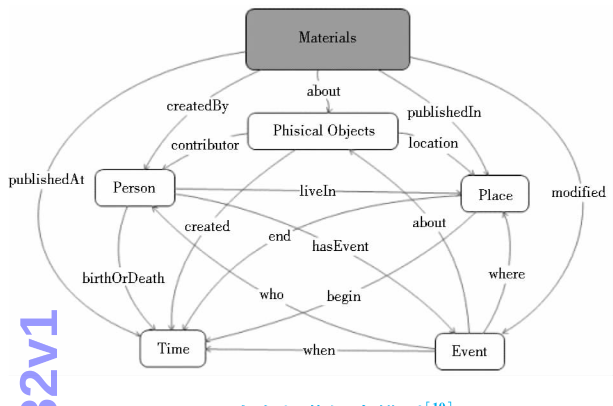


图 1　人名规范概念模型[10]

VIVO 是由美国国立卫生研究院(National Institutes of Health,NIH)资助的一个项目,由康奈尔大学、印第安纳大学和佛罗里达大学等 7 所学校和机构联合开发。该项目最初由康奈尔大学于 2003 年发起,目标是利用语义网及本体技术建设一个支持交流和发现的国家科学家网络。VIVO 利用关联数据、本体和可视化技术实现数字资源的语义互联,是国际上具有代表性的科研人员信息平台。VIVO 本体是科学家的数据模型,VIVO 本体中的主要的类有概念、时间、期刊、人员、出版、基金、机构等。VIVO 本体在设计时复用了其他 18 种标准词表。其中人员和机构复用了 FOAF,期刊复用了书目本体(Bibliographic Ontology,BIBO),基金由 VIVO 自定义的类和属性描述[12]。

此外,2010 年,匈牙利国家图书馆将书目数据和规范数据作为关联开放数据发布,人名规范数据采用 FOAF 描述[13]。2014 年,北卡罗莱纳州立大学图书馆将组织名称规范文档(The Organization Name Authority,ONA)发布为关联数据,其复用了 FOAF 描述机构[14]。

3.2　关联数据发布项目中的数据模型分析

数据建模首先是概念模型设计,其次是将概念模型形式化。概念模型设计指分析数据中的实体属性与实体间关系,抽象出实体对应的概念类、属性及其关系;将概念模型形式化是指采用规范的术语来描述抽象概念的属性、属性值及其关系。不同领域的社区已经开发或提出大量的元数据标准,这些标准因特定的目的而创建,包括以有效和一致的方式指导数据结构、数据值、数据内容和数据交换的设计、创建和实现。据此,元数据标准可分为 4 类:数据结构标准、数据内容标准、数据取值标准、数据交换标准。机器可读目录(Machine Readable Catalogue,MARC)或都柏林核心元素集(Dublin Core Element Set,DC)等属于结构标准,编目规则属于内容标准,分类法、主题词表、名称规范等属于取值标准,XML 等属于交换标准[15]。在本文关联数据模型构建中涉及的“标准词表”是数据结构标准,目的是定义数据的结构和语义。

在概念模型设计上,各项目根据自身数据的内容抽象实体和实体间关系建立了概念模型。各项目除了将规范对象建模为主要类,也根据需要将与之相关的信息建模为关联类。例如,上海图书馆将人员规范数据发布为关联数据,其提取了人员及与人员相关的描述项目作为概念类,而德国图书馆 GND 汇集了 GKD、PND、SWD 以及 EST 等多个规范文档,提取的类更为丰富。同时它们也存在差异:①各项目在部门相同描述项目的处理上存在差异,如上海图书馆将与人员相关联的时间设为类。德国国家图书馆 GND,将出生日期、出版日期等作为类的属性;②根据各自规范数据种类的复杂程度,一部分项目在大类下划分多个下层类,形成多层级类别体系,如上海图书馆、德国图书馆;一部分项目仅有单层类别体系,如 VIVO。在各类下都有揭示概念及概念间关系的详细元数据描述,将概念模型形式化。

各项目在模型形式化的处理上分为两类:①以一种标准词表为核心词表,以其他标准词表为补充达到对数据的准确描述。如 VIAF 以 Schema. org 为核心词表[5]。Schema. org[16] 由 Google, Microsoft, Yahoo 和 Yandex 赞助发起,它在广泛的范围内提供一致的语义描述词汇,并且允许根据应用需要在其结构下扩展。其词汇包括 600 多个类和 900 多种关系,覆盖范围包括个人、组织机构、地点、创造性作品、时间、医疗、商品等。OCLC 意识到 Schema. org 可能是有效描述图书馆信息资源的重要工具,它提供了在网络范围内整合 OCLC 丰富的书目数据、规范数据等信息资源的机会[17]。国内学者贾君枝认为 Schema. org 以其高通用性、高表达性以及丰富的数据类型具有描述中文人名规范数据的优势[18];②建立自定义的本体作为数据模型发布关联数据,在自定义本体中复用多种广泛使用的标准词表,或者与已有标准词表对齐,即多种标准词表的组合加自定义。如德国国家图书的数据模型是

GND 本体，上海图书馆的数据模型是 shlnames 本体，VIVO 项目的数据模型是 VIVO 本体[12,19]。这些项目中常复用的已有标准词表有 FOAF、vCard、BIBO 等（见表 2）。FOAF 是一个致力于用网络来联结人和信息的语义描述框架，为描述人员、团队、机构和文档提供了一系列类和属性。vCard 侧重描述人员和组织，包括位置信息等[20]。vCard 于 1995 年首次被提出，此后出现了新的标准词表 FOAF（2005）和 ORG 本体（2013）。FOAF 更多地关注人、代理、事物和社交网络实体之间的关系，而 ORG 本体更关注组织结构、角色和活动。3 种本体之间有一些重叠，它们可以分别提供有用的词汇，并且在协作使用时还可以提供增强的信息[21]。BIBO 是描述多种出版物及其相关资源的本体[22]。

表 2 关联数据发布项目的数据模型

序号	项目	国家	数据模型	主要实体大类	人员	机构	基金	期刊	地点	时间	事件
1	VIAF	美国	Schema.org 为核心	Place、Organization、Person、CreativeWork 等	schema.org	schema.org			schema.org		
2	德国国家图书馆 GND	德国	GND 本体	Person、Conference or event、Corporate body、Place or Geographic name、Subject Heading、Work	GND 自定义（FOAF）	GND 自定义（FOAF）			GND 自定义		Rdaregistry.info
3	上海图书馆人名规范库	中国	人名规范库本体（shlnames）	Person、Place、Time、Event、Materials、Phisical Objects	FOAF				PROV ontology、shl 自定义	Time	shl 自义定
4	VIVO	美国	VIVO 本体	Organization、Person、Concept、Journal、Grant、Date TimeValue and DateTimeInterval、Teaching 等	FOAF、vCard、VIVO 自定义	FOAF、vCard、VIVO 自定义	VIVO 自定义	BIBO	vCard	Vivo 自定义	

4 科研实体规范数据建模

结合已有关联数据项目的数据模型构建经验和 NSTL 名称规范库的数据特点，构建 NSTL 名称规范的关联数据模型。

4.1 概念模型设计

概念模型设计上，分析人员、机构、基金、期刊规范数据中的可抽象实体及实体间关系，建立抽象概念模型，见图 2。NSTL 名称规范数据的概念模型包含人员、机构、基金、作品、地点 5 个类及类之间的关系，“期刊”在概念模型中属于“作品”的下位类。

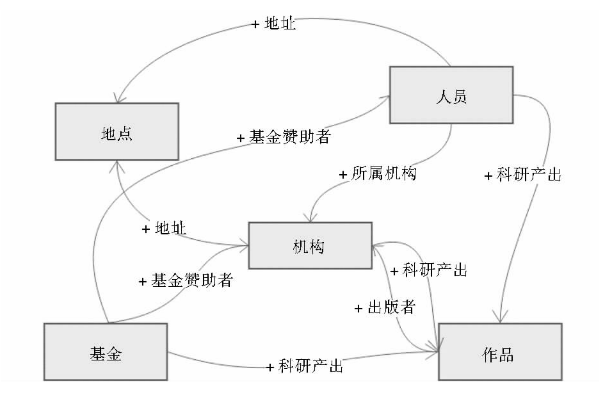


图 2 名称规范数据抽象模型

4.2 概念模型形式化

4.2.1 模型形式化原则

在模型形式化方面，W3C 的关联数据最佳实践建议在建模过程中应尽可能重用已有的标准词表，避免重新发明轮子，以促进关联数据的交换和共享[23]。已有的关联数据发布项目也遵循了这一原则，在构建关联数据模型时尽量复用已有标准词表或与标准词表对齐。现实中很难找到一种词表满足全部需求，可能需要复用多种标准词表。一般来说选择的词表种类越多，组合中的混合程度越高，关联数据模型也就更复杂，与外部数据关联难度越大。因此在确保标准词表质量的前提下，尽量降低复用标准词表的数量。本文在选择标准词表描述数据时遵循以下原则：①尽量复用关联数据发布项目中广泛使用的已有标准词表，保证较高的互操作性；②在满足描述需求的前提下，尽量采用使用广泛、维护良好的标准词表，使得关联数据模型更加易于理解；③尽量采用数量更少的标准词表组合，降低模型复杂度，便于与外部数据关联，降低维护成本；④对于多个标准词表组合仍不能覆盖的部分，根据需要按照自定义词表的最佳实践[24]定义符合需求的类和属性。

4.2.2 两种方案比较

如上文 3.2 节所述，已有的关联数据项目模型形式化方案主要有两类，究竟哪一类更加适合 NSTL 名称规范数据？本文设计两种关联数据模型形式化方案，从关联数据模型对 NSTL 名称规范数据的表达程度、模型复杂程度等角度进行比较，选取较优的方案。

参考关联数据发布项目的实践，采用项目中常用的标准词表，适当补充其他标准词表构建两种方案：

周毅，张建勇，刘峥，等. 科研实体名称规范的关联数据模型构建[J]. 图书情报工作,2020,64(10):109－117.

①以Schema. org 为核心词表方案,此方案参考 VIAF 以 Schema. org 为核心词表,以 FRAPO 配合描述基金;②多词表组合方案,该方案综合采用 FOAF、ORG、VIVO、BIBO、vCard 等组合描述,如表 3 所示:

表 3　以 Schema. org 为核心词表方案与多词表组合方案所采用的标准词表

	人员	机构	基金	期刊	地点
Schema. org 为核心词表方案	Schema. org	Schema. org	FRAPO	Schema. org	Schema. org
多词表组合方案	FOAF	ORG	VIVO	BIBO	vCard

分别使用两种方案与 NSTL 名称规范数据进行映射。NSTL 名称规范数据采用元素和属性组合描述,如日期由元素“date”描述,要限定日期的类型,需要同时采用属性“日期类型(date-type)”。“出生日期”表示为 <date date-type = “birth” > …… </date >。在构建关联数据模型时要同时考虑元数据的元素和属性。

在映射过程中,有 3 种不同的映射方式:①一对一映射,即一个元素映射一个属性。如元素 given-names,对应 schema. org 中的 schema:givenName;②一对多映射,即一个元素因属性限定不同,可能映射到关联数据模型的多个属性。如日期因其属性值不同可以分为出生日期和死亡日期,出生日期 <date date-type = “birth” >…… </date > 映射 schema:birthDate,死亡日期 <date date-type = “death” > …… </date > 映射 schema:deathDate;③多对一映射,如规范数据的日期由元素年(year)、月(month)、日(day)组合描述。例如出版日期是 2019 年 1 月 1 日表示为: <date date-type = “pub-date” > <day >01 </day > <month >01 </month > <year >2019 </year > </date >,映射到 Schema. org 的 schema:datePublished。

映射结果显示两种方案均有不能表达的 NSTL 名称规范元素和属性。统计两者不能表达的元素和属性数量,并进行比较。以机构为例,元素映射情况见表 4,属性映射情况见表 5。在机构描述上,Schema. org 描述机构的属性丰富,仅个别元素和属性不能表达;相比之下 ORG 在表达机构规范数据时,缺少表达日期、外部链接的属性。此外,人员方面,Schema. org 也具备丰富的人员属性,仅缺少学位、专业方面的少量属性。FOAF 关于人员的职业、教育信息的属性较少,总体上 Schema. org 的表达程度更高;在基金描述上,与 VIVO 相比,FRAPO 描述基金项目的货币类型、主题、关键词等的属性丰富;期刊描述上,BIBO 与 Schema. org 相当,Schema. org 的标题类别相对丰富,二者都未对期的类别,如总期、分期、增期进行区分;地址描述上,Schema. org 与 vCard 描述一致。

表 4　两种方案未能表达的机构元素比较

序号	NSTL 名称规范元素中文名称	NSTL 名称规范元素名称	Schema. org 为核心词表方案	多词表组合方案
1	研究方向	research-subject	X	X
2	外部链接	ext-link		X
3	注释	notes		X
4	日期	date		X
5	日	day		X
6	月份	month		X
7	季度	season		X
8	年份	year		X

注:表中“X”表示没有相应的映射项

表 5　两种方案未能表达的机构属性比较

序号	NSTL 名称规范属性中文名称	NSTL 名称规范属性名称	Schema. org 为核心词表方案	多词表组合方案
1	语种	xml:lang	X	X
2	外部链接类型	ext-link-type		X
3	超链接	xlink:href		X
4	注释类型	notes-type		X
5	日历类型	calendar	X	X
6	日期类型	date-type		X
7	GB/T 7408 格式日期	gbt-7408-date	X	X

分类统计两种方案未能表达的元素和属性,结果见图 3。左边柱形表示 Schema. org 为核心词表方案未能表达的元素和属性,右边表示多词表组合方案。总体上来说,多词表组合方案未能表达的名称规范数据元素和属性均高于 Schema. org 为核心词表方案。因此在对规范数据的表达完整程度上,以 Schema. org 为核心词表方案优于多词表组合方案。

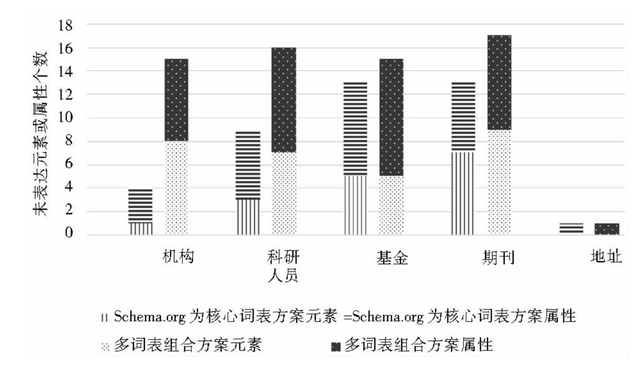


图 3　Schema. org 为核心词表方案、多词表组合方案未能表达的规范数据的元素和属性统计

另外在模型复杂度上,Schema. org 为核心词表方案以 Schema. org 为核心标准词表配合基金标准词表 FRAPO,相比于由 4 种标准词表混合的多词表组合方

案来说,模型复杂度更低,更易与外部数据关联,也更易于更新和维护。

词表使用上,以 Schema. org 为核心词表方案中,作为核心词表的 Schema. org[16] 由几大搜索引擎开发,对搜索引擎具有天然的友好性,使名称规范数据更易于被网络发现。Schema. org 本身具有丰富的数据类型,基本能准确地描述名称规范数据中的人员、机构、期刊、地点,采用同一标准词表描述多种对象,有利于保持描述的一致性,同时更易于与 VIAF 等已经以 Schema. org 为核心词表的关联数据集关联。

因此,NSTL 名称规范系统的关联数据模型采用 Schema. org 为核心词表方案更优。

4.3 NSTL 名称规范数据的关联数据模型

采用以 Schema. org 为核心词表的方案构建 NSTL 名称规范数据的关联数据模型,如图 4 所示(篇幅所限仅显示部分属性),概念模型中的人员、机构、基金、作品、地址分别表示为:Schema:Person、Schema:Organization、Frapo:Funding、Schema:CreativeWork、Schema:postalAddress。Schema:CreativeWork 表示所有类型的作品,期刊表示为 Schema:Periodal,是 Schema:CreativeWork 的子类。

各类之间的实线连线揭示 NSTL 名称规范数据支持的类间关系。如人员与机构间的隶属关系(Schema:affiliation)、机构与地址间的位置关系(schema:location)。作品与人员、机构、基金的连线,将作者、机构、基金和他们的产出相关联。

与类以虚线相连的是数据属性,各类所表示的实体都有一个唯一标识的 URI,其他类型的标识符也可以由属性"Schema:identifier"表示,各类的名称由"Schema:name""schema:alternateName"表示,人员的名称属性更为丰富。

此外,Schema. org 中的"Schema:sameAs"可以将本地的规范数据与外部的数据集连接,如 DBpedia、VIAF。

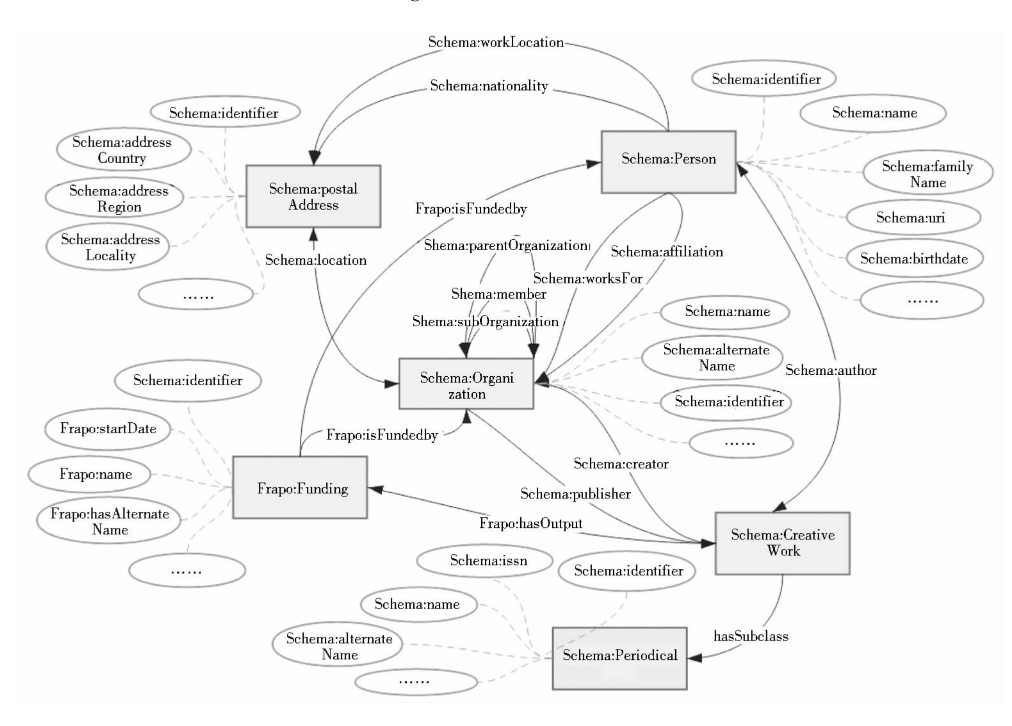


图 4 NSTL 名称规范关联数据模型

5 科研实体关联数据发布实验

从 NSTL 名称规范系统中选取样例数据,包括人员、机构、基金、地址信息各 2 000 条。采用目前广为使用的 D2RQ 开源软件进行名称规范数据的发布实验。

D2R[25] 是德国柏林自由大学大学的开源项目,主要功能是将关系型数据库中的数据转换成动态虚拟的资源描述框架(Resource Description Framework,RDF)数据。同时以 RDF 视图的方式,实现对关系型数据库中的数据进行查询访问。

D2RQ 映射语言提供了 URI 生成的描述模板机制,用来指导生成实例化资源的真实 URI。本次采用绝对 URI 模式,即使用 d2rq:uriPattern 属性对每个表

对象进行配置，形如 JournalPaper/@@JournalPaper.PAPER_ID@@，“/”前面的部分为表名，“@@”之间的部分为表的列。

使用 D2RQ Mapping 映射语言，基于 4.3 节中 NSTL 名称规范数据的关联数据模型，完成几类规范对象数据从关系数据表向类、属性及关系的语义映射，创建现有规范数据映射到关联数据模型的映射文件，如图 5 所示：

```
# Table organization
map:organization a d2rq:ClassMap;
    d2rq:dataStorage map:database;
    d2rq:uriPattern "organization/@@organization.institution-id|urli:
    d2rq:class schema:organization;
    d2rq:classDefinitionLabel "organization";
    .
map:organization__label a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:organization;
    d2rq:property rdfs:label;
    d2rq:pattern "organization #@@organization.institution-id@@";
    .
map:organization_institution-id a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:organization;
    d2rq:property schema:identifier;
    d2rq:propertyDefinitionLabel "organization institution-id";
    d2rq:column "organization.institution-id";
```

图 5　映射文件

通过映射文件对关系型数据库的数据进行转换和访问，将名称规范数据发布为关联数据。发布为关联数据后机构的详情页见图 6，以北京大学为例，点击图中表示地点的字段 schema:location 即可链接到相应的地点，实现实体间的互联，见图 7。

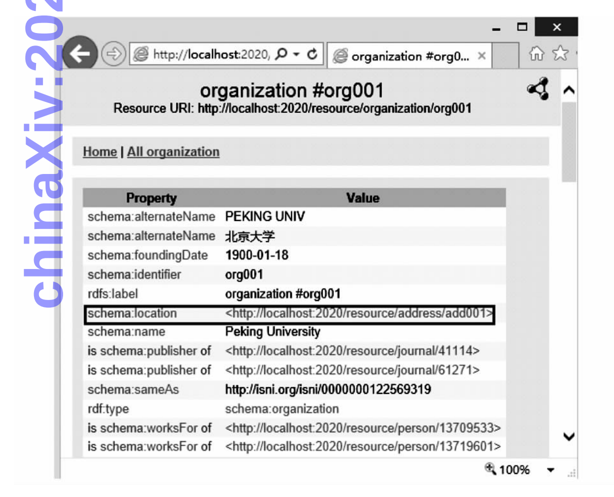


图 6　机构详情－北京大学

6　讨论与结论

通过实验将 NSTL 名称规范数据成功发布成为关联数据，展示了关联数据状态下的规范对象及其关联。结合 4.2.2 中以 Schema.org 为核心词表方案与 NSTL 名称规范数据元数据的映射情况，在将 NSTL 名称规范数据发布为关联数据的过程中也发现一些问题：①没有通用数据模型能够完全表达所有数据。尽管在构建数据模型时尽可能地复用已有的标准词表，以方便



图 7　地址详情－北京大学地址

与外部数据集的连接，但仍有部分数据不能在已有标准词表中找到相应表达（见表 4、表 5），需要自定义。自定义类或属性又增加了数据模型的复杂性，不利于数据的复用和互联；②在数据转换为关联数据的过程中可能造成数据损失。由于关联数据模型与原有数据模型在描述粒度上的不一致，原有数据模型中更细粒度的数据可能丢失，如 NSTL 名称规范数据的元数据中，元素“来源题名缩写（abbrev-source-title）”有元素属性“abbrev-type”对不同的缩写类型进行区分，元素“日期（date）”有元素属性“calendar”对日期所属的不同日历类型进行区分，目前用于关联数据发布的标准词表中没有同等粒度的描述。因此，将在 NSTL 规范数据发布为关联数据时，可能存在数据损失。但这一问题可通过将数据中的日历类型或缩写类型统一为一

种的方法进行处理。

综上所述,本文在分析国内外关联数据发布项目的基础上,构建 NSTL 名称规范数据的关联数据模型。过程中比较了多种标准词表组合和以一种标准词表为核心的两种方案,结果显示以 Schema. org 为核心词表的数据模型相对更加能够准确地表达名称规范数据,同时更易于融入互联网。因此,选择以 Schema. org 为核心词表的方案作为 NSTL 名称规范关联数据模型。最后采用 D2RQ 选取样例数据进行了关联数据的发布实验,呈现了名称规范数据发布为关联数据后的结果,以期为以后其他机构的关联规范数据建模提供参考。

研究规范数据的关联数据模型,将 NSTL 的名称规范数据发布为关联数据,有助于数据的共享、互联、质量提升,便于融入互联网,同时也为其他机构重用、发布关联数据提供模型参考。未来需要研究规范数据发布为关联数据后的应用价值和前景,研究如何实现更大范围内的数据关联和自身质量提升,以及研究如何利用关联数据化的规范数据构建更加丰富的知识服务。

参考文献:

[1] 曾建勋. 基于海量数字资源的科研关系网络构建探究[J]. 情报学报,2013,32(9):929-935.
[2] 刘炜, 张春景, 夏翠娟. 万维网时代的规范控制[J]. 中国图书馆学报, 2015(3):2-33.
[3] VIAF [EB/OL]. [2019-06-25]. https://www.oclc.org/zh-Hans/viaf.html.
[4] THOMAS B H, JEFFREY A Y. Description of the VIAF (virtual international authority file) dataset[EB/OL]. [2019-06-25]. http://www.semantic-web-journal.net/sites/default/files/swj294.pdf.
[5] VIAF. RDF changes [EB/OL]. [2019-06-25]. https://outgoing.typepad.com/outgoing/2015/04/viaf-rdf-changes.html.
[6] OCLC adds linked data to WorldCat.org[EB/OL]. [2019-06-25]. http://www.oclc.org/news/releases/2012/201238.en.html.
[7] Versionshistorie des linked-data-servicE[EB/OL]. [2019-11-22]. https://www.dnb.de/DE/Professionell/Metadatendienste/Datenbezug/LDS/lds_versionshistorie.html? nn=250612.
[8] GND ontology[EB/OL]. [2019-06-25]. https://d-nb.info/standards/elementset/gnd#CharactersOrMorphemes.
[9] Gemeinsame normdatei (GND)[EB/OL]. [2019-06-25]. https://lod-cloud.net/dataset/dnb-gemeinsame-normdatei.
[10] XIA C J, LIU W. Name authority control in digital humanities: building a name authority database of Shanghai Library[J]. International journal of libraryship 2018,3(1): 21-35.
[11] 上海图书馆. 人名规范库本体(shlnames)[EB/OL]. [2019-06-25]. http://data.library.sh.cn/ont/ontology/tree? g=http://ont.library.sh.cn/graph/shlnames.
[12] WOODS A. Source ontologies for VIVO[EB/OL]. [2019-06-25]. https://wiki.duraspace.org/display/VIVODOC110x/Source+ontologies+for+VIVO.
[13] NATIONAL LIBRARY OF HUNGARY. National széchényi library (national library of Hungary) on the semantic Web[EB/OL]. [2019-06-25]. http://nektar.oszk.hu/wiki/Semantic_web.
[14] HANSON E M. A beginner"s guide to creating library linked data: lessons from ncsu\"s organization name linked data project [J]. Serials review,2014,40(4):251-258.
[15] 胡小菁. 文献编目:从数字化到数据化[J]. 中国图书馆学报, 2019, 45(3):49-61.
[16] GOOGLE, YAHOO, MICROSOFT, et al. Schema.org[EB/OL]. [2019-10-10]. https://schema.org/.
[17] 张雪松, 谈海蓉, 姚湘中. 网络书目资源描述规范 SchemaBibEx 及其应用[J]. 图书馆杂志, 2016 (5):67-75.
[18] 贾君枝, 石燕青. 中文个人名称规范文档的关联数据化研究[J]. 情报学报, 2016, 35(7):696-703.
[19] 李柏炀. 基于关联数据的科研关系揭示研究[D]. 长春:东北师范大学,2016.
[20] NOGALES A M, ELENA G B. Measuring vocabulary use in the linked data cloud[J]. Online information review,2017,41(2):252-271.
[21] IANNELLA R, MCKINNEY J, vCard ontology - for describing people and organizations [EB/OL]. [2019-06-25]. https://www.w3.org/TR/vcard-rdf/.
[22] D'ARCUS B, GIASSON F. Bibliographic ontology specification [EB/OL]. [2019-06-25]. http://bibliontology.com/.
[23] HYLAND B, ATEMEZING G, TERRAZAS V B. Best practices for publishing linked data[EB/OL]. [2019-05-23]. http://www.w3.org/TR/ld-bp/.
[24] 伍德,扎伊德曼,鲁思,等. 关联数据:万维网上的结构化数据[M]. 蒋楠,译. 北京:人民邮电出版社,2018:3.
[25] BIZER C, CYGANIAK R. D2R server-publishing relational databases on the semantic [EB/OL]. [2019-07-13]. http://richard.cyganiak.de/2008/papers/d2r-server-iswc2006.pdf.

作者贡献说明:

周毅:撰写并修改论文;
张建勇:负责 NSTL 名称规范数据及实验;
刘峥:搭建论文框架,修改论文;
刘秀敏:提供发布实验技术支持。

Research on the Construction of Linked Data Model for Research Entity's Name Authority Data

Zhou Yi　Zhang Jianyong　Liu Zheng　Liu Xiumin
National Science Library, Chinese Academy of Sciences, Beijing 100190

Abstract: [Purpose/significance] The purpose of this paper is to study the linked data model of publishing the NSTL's research entity name authority data as linked data. After the name authority data is published as linked data, it can be reused as an open linked data set by other system or organization, and also can be better integrated with other linked data sets to improve data quality. In addition, it also provides a model building reference for other organizations to publish authority data as linked data. [Method/process] First, this paper analyzed and compared the data models used in the linked data publishing projects at home and abroad. It showed that the data models in the linked data publishing projects were mainly divided into two categories. Then, combined with the characteristics of NSTL name authority data, two forms of linked data models were designed. It compared the two models from the expression level of the NSTL's data and the complexity of the models. The better one was selected. Finally, it used D2RQ as tool to publish the sample data as linked data. [Result/conclusion] The analysis found that the model with Schema.org as the core standard vocabulary has better performance. So it is more suitable as a linked data model for NSTL's name authority data.

Keywords: research entities　name authority　linked data　data model